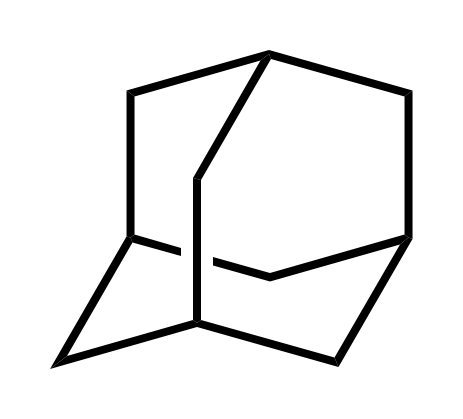
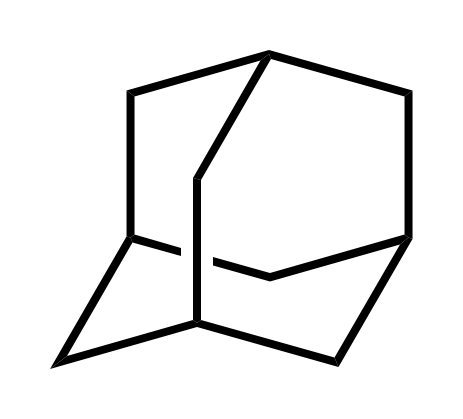
Sheryl Oilar

I571 Homework 1

Due 9/23/13



[](http://upload.wikimedia.org/wikipedia/commons/5/58/Adamantane_acsv.svg)C1C2CC3CC1CC(C3)C2

[](http://upload.wikimedia.org/wikipedia/commons/5/58/Adamantane_acsv.svg) C1C2CC3CC(CC1C3)C2

1. Hydroxypyridine’s can change structure based on the environment they are in. In an acidic environment the oxygen will likely be bonded to H and the ring will be aromatic. In a basic environment, the oxygen will be double bonded to the ring and the ring will no longer be aromatic. This means two very different strauctures can be used to represent the structure.

For example

cccnccO CCCNCC=O

1. A molecule can be represented using connection tables or SMILES strings in N! ways for a molecule with N atoms, making it unfeasible to efficiently identify molecules. A canonical representation helps solve this by giving a unique ordering to atoms based on the Morgan algorithm in a molecular graph using the following steps.
   1. First – Each atom is assigned a connectivity value based on the number of bonded atoms
   2. Second – New connectivity value assigned based on the sum of the connectivity values of the atoms neighbors
   3. Second step is repeated until the molecule has the maximum number of unique connectivity values.
   4. Atom with the highest connectivity value is the first atom in the connection table
   5. If there is a tie, other properties are considered such as atomic number and bond order.    

Final connectivity values = 

1. A **Structural Key** uses a bit string, string of 0’s and 1’s based on the presence or absence of a specific component such as substructures (rings, groups) or number of features (groups of atoms or substructures) contained in each molecule to quickly describe molecules, making them easier to query. Each molecule is identified using a fragment dictionary. Each fragment is in a different spot. When a new molecule is added to the database, it is compared against the fragment dictionary. A 0 or 1 is assigned to each I’th spot based on the presence or absence of each corresponding feature in the fragment dictionary. This database is used to quickly identify or eliminate molecules in a substructure search. MACCS and ISIS use structural key and fragment dictionary method.

A **Hashed Fingerprint** does not require a predefined fragment dictionary making it applicable to searching molecules in any database and from anywhere because they are produced by generating all possible linear paths through a molecule. These paths are used within a hashing program to assign 5-6 bits which make the fingerprint of the molecule. Bits may be assigned to more than 1 pattern leading to potential false negatives when searching in this method.

The structural key, fragment library method would be preferable because this method results in fewer false negatives and is much faster at eliminating compounds that do not immediately meet the criteria of the substructure search. It is faster and less likely to miss compounds.

1. A Tanimoto coefficient of 0.89 (scale 0-1) indicates the two compounds are very structurally similar in regards to molecular features and they likely could have very similar biological properties. The Tanimoto coefficient indicates they have similar fingerprints but may not be identical molecules. The similar fingerprints indicates the second molecule may be worth further investigation but might not have the same biological function.
   1. Using Open Babel

|  |  |  |
| --- | --- | --- |
| SMARTS | Excluded SMARTS | Frequency |
| O=N(~O)a | O=N(O)c[$(aS(=O)=O),$(aaS(=O)=O),$(aaaS(=O)=O),$(aC((F)F)F),$(aaC((F)F)F),$(aaaC((F)F)F)] | 330 |
| a[NH2] | [NH2]a[$(C((F)F)F),$(S(=O)=O),$(C(=O)O)],$(aa[$(C((F)F)F),$(S(=O)=O),$(C(=O)O)],$(aaa[$(C((F)F)F),$(=O)=O),$(C(=O)O)])] | 131 |
| a[N;X2]=O |  | 0 |
| CO[N;X2]=O |  | 0 |
| N[N;X2]=O |  | 1 |
| O1[c,C]-[c,C]1 |  | 11 |
| C1NC1 |  | 2 |
| N=[N+]=[N-] |  | 0 |
| C=[N+]=[N-] |  | 0 |
| N=N-N |  | 1 |
| c[N;X2]!@;=[N;X2]c | [$([N;X2]($(acS((=O)=O))),$(aacS((=O)=O)),$(aaacS((=O)=O)),$(aaaacS((=O)=O))])=[N;X2][$(acS((=O)=O)),$(aacS((=O)=O)),$(aaacS((=O)=O)),$(aaaacS((=O)=O))])] | 5 |
| cN!@;=[N;X3](O)c |  | 0 |
| [OH,NH2][N,O] | O=N(O)[O-] | 17 |
| [OH]O |  | 0 |
| [OH][N;X2]=C |  | 4 |
| [c,C]OO[c,C] |  | 5 |
| C[NH][NH]C |  | 21 |
| [OH]Na |  | 0 |
| [OH]N | [$([OH]Na),$([OH][N;X2]=\*)] | 11 |
| [NH2]Na |  | 2 |
| [NH2]N | [NH2]Na | 6 |
| [OH][N;X2]=[N;X2] |  | 0 |
| [CL,Br,I]C |  | 0 |
| [Cl,Br,I]C=O |  | 0 |
| [N,S]!@[C;X4]!@[CH2][Cl,Br,I] |  | 1 |
| [Cl,Br,I][C;X4] | [$([Cl,Br,I][C;X4][F,Cl,Br,I]),$([Cl,Br,I]C((C)C)] | 52 |
| SC[Cl] |  | 3 |
| [Cl,Br,I]!@[C;X4!@[C;X4]O |  | 0 |
| [Cl]C([X1])=C[X1] |  | 2 |
| [Cl,Br,I][CH][CH3] | [Cl,Br,I][C]([Cl,Br,I,F])[CH3] | 1 |
| [Cl,Br,I]C((F,Cl,Br,I])[X1])C=C |  | 0 |
| [Cl,Br,I]C((F,Cl,Br,I])[X1])C(=O)[c,C] |  | 0 |
| [cH]1[cH]ccc2c1c3c(cc2)cc[cH]cH]3 |  | 0 |
| [Ch]1cccc2c1[cH][cH]c3c2ccc[cH]3 |  | 0 |

Most of the SMARTS strings that allowed for substitution showed higher frequency counts than those that were more rigid or contained less common atoms